A Statistical Prescription to Estimate Properly Normalized Distributions of Different Particle Species

M. Casarsa[a1] P. Catastini[b2] G. Punzi[c3] L. Ristori[d4]

[a] *Fermi National Accelerator Laboratory, Batavia, USA*
[b] *Fermi National Accelerator Laboratory, Batavia, USA*
[c] *Università di Pisa and INFN Sez. Pisa, Pisa, Italy*
[d] *INFN Sez. Pisa, Pisa, Italy*

Abstract

We describe a statistical method to avoid biased estimation of the content of different particle species. We consider the case when the particle identification information strongly depends on some kinematical variables, whose distributions are unknown and different for each particles species. We show that the proposed procedure provides properly normalized and completely data-driven estimation of the unknown distributions without any a priori assumption on their functional form. Moreover, we demonstrate that the method can be generalized to any kinematical distribution of the particles.

1 Introduction

The estimation of the particle species content in a sample of reconstructed tracks is a recurrent problem in Particle Physics. To this purpose, different experimental techniques are used to obtain information about the particle type; typical examples are the measurement of the particle Time-of-Flight (ToF) from the production vertex to a given position inside the detector, or the measurement of the particle energy loss per unit length of the traveled path due to the interaction with the detector material (dE/dx).

The information provided by these techniques is related to the particle type but typically depends also on the track momentum.

It is common practice to include the particle identification (PID) information in a Maximum Likelihood (ML) fit in order to estimate the particle species content of the sample. On the other hand, the strong momentum dependence of the separation power between different particles may lead to strongly biased results, if not properly treated in the ML fit.

To be more specific, let's consider a mixture of known particle species, for example pions

[1] casarsa@fnal.gov
[2] pierluigi.catastini@pi.infn.it
[3] giovanni.punzi@pi.infn.it
[4] luciano@fnal.gov

(π), kaons (K), protons (p), and electrons (e), and assume that the *PID* information is provided by the dE/dx measurement in a drift chamber. Using the separating power provided by the *PID*, we want to estimate the unknown fractions f_π, f_K, f_p, and f_e of each particle type contained in the sample by means of a ML fit. Our observables are the measured dE/dx response (that we will indicate as x) and the momentum of the particle p. We then label as t_j the particle hypothesis and the conditional probability density function of x_i for track i, given p_i and t_j, will be indicated as $\mathcal{P}(x_i|p_i,\, t_j)$.
Finally, the likelihood function is expressed as:

$$\begin{aligned} L(f_j) &= \prod_i \Big(\sum_{j=\pi,K,p,e} f_j \mathcal{P}(x_i,\, p_i \,|\, t_j)\Big) \\ &= \prod_i \Big(\sum_{j=\pi,K,p,e} f_j \mathcal{P}(x_i|p_i,\, t_j) \times \mathcal{P}(p_i|t_j) \quad \Big) \ , \end{aligned} \tag{1}$$

where i is the track index, with the additional condition:

$$\sum_{j=\pi,K,p,e} f_j = 1 \ . \tag{2}$$

In practice, we often have poor information on the distributions of the additional observables ($\mathcal{P}(p_i|t_j)$ in our example). Sometimes they are completely unknown. This is the case, for example, of particles produced during the hadronization of B mesons[5]: the momentum distribution of each particle type is unknown and the correct likelihood function as defined in (1) cannot be constructed.
Note that using the conditional likelihood function:

$$L(f_j) = \prod_i \Big(\sum_{j=\pi,K,p,e} f_j \mathcal{P}(x_i|p_i,\, t_j)\Big) \tag{3}$$

may lead to strongly biased results, if our additional variable, the momentum in our example, has different distributions for different particle types since the $\mathcal{P}(p_i|t_j)$ term cannot be factorized in (1). As discussed in [2] and shown in [3] specifically for the particle species estimation, whenever the templates used in a multi-component fit depend on additional observables, it is necessary to use the complete likelihood expression, and explicitly include the probability distributions of all observables. In our example this implies that we need to include the momentum distributions of each particle type in the likelihood. The crucial question is how to avoid strong bias in the particle fraction estimation when the momentum distributions of each particle type are unknown.
In [3] it was shown that a possible solution to this problem is to use a series expansion of the unknown distributions; the Fourier coefficients of the series are free parameters determined by the fit.

[5] This work was motivated by the effort of understanding the properties of particles produced during B mesons hadronization, that represents a major issue in the development of flavor tagging algorithms like the Same Side Kaon Tagging [1].

In this paper we propose a different strategy, based on the idea that if the fit is performed in sufficiently small momentum intervals, the bias due to the use of the likelihood function (3) is small and it goes to zero as the momentum interval width decreases. In Sec. 2 we present a simple procedure to perform the fit and show that we can extract the unknown momentum distributions of each particle type in a completely data driven mode without any *a priori* assumption on the corresponding functional forms.
In Sec. 3, given the results of our fitting procedure, we describe a powerful technique to extract the distribution of any kinematical variable for any particle type, with the proper normalization. Finally, in Sec. 4 some applications of the method to Pythia [4] Monte Carlo samples simulated with the CDF detector [5] are shown.

2 Estimation of Particle Fractions and Momentum Distributions

As anticipated above, we restrict ourselves to the particles contained in small momentum intervals of equal width Δp. In each bin we estimate the particle content using a ML fit by observing that if Δp is sufficiently small, the bias introduced by using the conditional likelihood of (3) is negligible. As will be evident in the following, we find more convenient to use an Extended Likelihood (EL) fit (see [6] for a definition). In each momentum bin m the Extended Likelihood function takes the form:

$$\begin{aligned}\log(L_m) = \mathcal{L}_m &= \sum_{i=0}^{N_m} \log(\sum_{j=1}^{M} N_{j,m}\, \mathcal{P}_j(x_i|m,t_j)) - N_m \\ &= \sum_{i=0}^{N_m} \log(\sum_{j=1}^{M} N_{j,m}\, \mathcal{P}_j(x_i|m,t_j)) - \sum_{j=1}^{M} N_{j,m}\ ,\end{aligned} \tag{4}$$

where N_m is the total number of particles in the m-th momentum bin, M is the number of different particle species, $N_{j,m}$ is the number of particles of type j, x_i is the *PID* response for the m-th momentum bin and $\mathcal{P}_j(x_i|m,t_j)$ is the conditional probability density function associated to x for the particle type t_j in the m-th momentum bin . In the rest of the paper we will simplify the notation by setting $\mathcal{P}_j(x_i) = \mathcal{P}_j(x_i|m,t_j)$. The EL must be maximized with respect to the free parameters $N_{j,m}$.
We notice that, given its particular form, the first derivative of our EL function can be evaluated analytically as:

$$\frac{\partial \mathcal{L}_m}{\partial N_{j,m}} = \sum_{i=1}^{N_m} \frac{\mathcal{P}_j(x_i)}{\sum\limits_{k=1}^{M} N_{k,m}\mathcal{P}_k(x_i)} - 1\ . \tag{5}$$

The critical points of (5) can be obtained using an iterative algorithm. In particular, starting from a first guess on our parameters, $N^0_{j,m}$, we can define an iterative procedure, a *Picard iteration* [7], of the form:

$$N_{j,m}^{n} = \sum_{i=1}^{N_m} \frac{N_{j,m}^{n-1} \mathcal{P}_j(x_i)}{\sum\limits_{k=1}^{M} N_{k,m}^{n-1} \mathcal{P}_k(x_i)} \ , \tag{6}$$

where at every step n of the iteration, we estimate the N_j^n that will be used as input values for the step $n+1$. The iteration converges to the roots of (5). Similarly, we can write the analytical expression of the second derivative of (5) and estimate the statistical error of our parameters from the inverse of the covariance matrix $V_{jl,m}^{-1}$ of the fit in the m-th momentum bin as:

$$V_{jl,m}^{-1} = \frac{\partial(-\mathcal{L}_m)}{\partial N_{j,m} \partial N_{l,m}} = \sum_{i=0}^{N_m} \frac{\mathcal{P}_j(x_i)\mathcal{P}_l(x_i)}{\left(\sum\limits_{k=1}^{M} N_{k,m} \mathcal{P}_k(x_i) \right)^2} \ . \tag{7}$$

Our iterative procedure is equivalent to the *Channel Likelihood* method introduced in [8]. It can be easily shown that the Channel Likelihood method is a maximization of an EL function of the form (5) using an iterative scheme similar to the one we propose. This approach has several advantages: no functional shape is assumed for the momentum spectra, the method, provided the $\mathcal{P}_j(x)$ templates, is completely data driven; all the fits can be performed in parallel, that is the iteration can be done in all momentum bins at the same time; the algorithm is very fast and stable respect to the initial guess on the $N_{j,m}$.

Once we obtain convergence of the iterative process, we can write the fraction of each particle type as:

$$f_j = \frac{N_j}{N} = \frac{1}{N} \sum_m N_{j,m} \ . \tag{8}$$

Observing that each bin content is fitted independently, the corresponding statistical uncertainty is:

$$\sigma(f_j) = \frac{\sqrt{\sigma^2(N_j)}}{N} = \frac{1}{N} \sqrt{\sum_m \sigma^2(N_{j,m})} \ . \tag{9}$$

Finally, since we are performing the fit in each momentum bin, the arrays of $N_{j,m}$ can be interpreted as *histograms*, that reasonably approximate the true momentum distributions of each particle type. In this way, we obtained an unbiased estimation of the particle composition, although the $\mathcal{P}(p_i|t_j)$ were unknown, and, at the same time, a reasonable estimation of the $\mathcal{P}(p_i|t_j)$ themselves.

We test our method on parametric Monte Carlo samples. Each sample consists of 20,000 particles. Different momentum spectra are used to generate each particle type and the corresponding fractions are fixed at: $f_\pi = 0.74$, $f_K = 0.17$, $f_P = 0.07$, and $f_e = 0.02$. We divide the sample in 50 momentum bins from 0.45 GeV/c up to 5 GeV/c and a EL fit is performed in each momentum bin. If no particle falls in a given

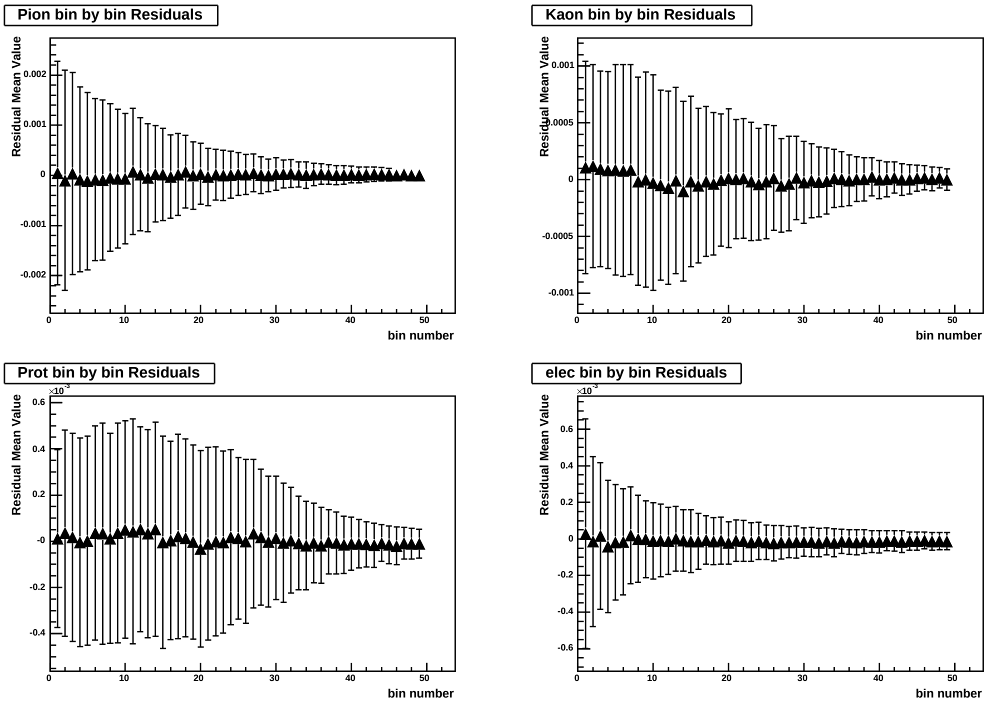

Figure 1: Mean and width of the distribution of the difference of the true and measured fractions in each momentum bin.

bin, the corresponding fit is not performed for obvious reasons. We repeat the fit on 500 parametric samples to extract the residual distributions of the estimators. No significant bias is observed as summarized in Tab. 1.

Another interesting test is the residual distribution of the estimators in each momentum bin for the 500 samples. Fig. 1 shows the residuals of the four estimators as a function of the bin number (i.e. momentum). In each plot, each point represents the mean value, while the error bar represents the width of the distribution of the residual. We observe no significant bias. Examples of the unknown momentum distributions obtained from the fit are shown in Fig. 2.

A very good agreement between the true distributions (filled histograms) and the fit (dots) is observed for all particle types.

	$f_\pi^{true} - f_\pi^{est.}$	$f_k^{true} - f_k^{est.}$	$f_p^{true} - f_p^{est.}$	$f_e^{true} - f_e^{est.}$
EL	$(0.2 \pm 2) \times 10^{-4}$	$(2 \pm 2) \times 10^{-4}$	$(3 \pm 1) \times 10^{-4}$	$(5 \pm 1) \times 10^{-4}$

Table 1: Mean values of the residuals of the particle fractions (integrated on all momentum bins) estimated in 500 parametric samples.

3 Extracting distributions of other quantities

Besides estimating the momentum spectra for the different particle species, we are typically also interested in obtaining distributions of additional relevant kinematical variables. We achieve this by combining our fitting procedure with a technique known as *sPlot* [9], which allows to estimate the composition of a mixture of several components and their covariance matrix by means of an Extended Maximum-Likelihood fit. The main requirement for the *sPlot* technique is that the kinematical variables to be plotted be uncorrelated with the discriminating variables of the fit. The output is represented by an histogram, called *sPlot*, of the relevant variables.

In our problem, this means that the variables we want to plot must not be correlated with the *pdf*'s that describe the *PID* response. However, any dependence of the *pdf*'s on kinematical variables is *frozen* by our strategy to perform a separate fit in each momentum bin: the correlation of any variable with the momentum, if any, can be neglected inside a sufficiently small bin.

For each momentum bin, our fit provides an estimate of the particle content and the corresponding covariance matrix. It follows that for each bin we can produce the *sPlot* of any kinematical variable.

Suppose we want to produce the $sPlot_{j,m}(y)$ of a given variable y for the particle type j, corresponding to the EL fit performed in the m-th bin. As explained in [9], we have to properly combine the result of the fit in the m-th bin with the corresponding covariance matrix to define an event by event weight called the s-weight, $sw_{ij,m}(x)$. We then have to fill an histogram of y where each event i, falling in the m-th bin, is weighted according to the s-weight:

$$sw_{ij,m}(x) = \frac{\sum\limits_{l=1}^{M} V_{jl,m}\,\mathcal{P}_l(x_i)}{\sum\limits_{k=1}^{M} N_{k,m}\,\mathcal{P}_k(x_i)} \; , \tag{10}$$

where $N_{k,m}$ are obtained by performing the iteration (6) and $V_{jl,m}$ is the covariance matrix (7) corresponding to the fit performed in the m-th bin. Applying the above procedure in each momentum interval, we obtain an array of $sPlot_{j,m}(y)$ histograms. We then exploit the additive property of the *sPlot*'s to add together all the $sPlot_{j,m}(y)$ corresponding to a given variable, one for each momentum bin; the resulting *sPlot* represents the distribution of the given variable y for the whole momentum range:

$$sPlot_j(y) = \sum_m sPlot_{j,m}(y) \; . \tag{11}$$

Thanks to the combination of the *sPlot* and our strategy of performing several fits in small momentum bins, we are able to obtain the distribution of any kinematical variable of each particle type belonging to our initial sample.

The method can be summarized in a three-step procedure as follows:

1. Fit the particle species content in several momentum intervals by means of the Extended Likelihood method via an iterative scheme.

2. In each momentum bin, using the results of the fits, produce the *sPlot*'s of any additional variable you are interested in for all the particle species.

3. Separately for each particle type, add the array of *sPlot*'s of a given variable to extract the distribution of the variable on the whole data sample.

4 Monte Carlo Study of the Complete Procedure

To test the whole procedure in a realistic case, we use a Pythia Monte Carlo [4] sample of $B^- \to D^0\pi^-$ events processed with the CDF detector simulation plus a parametric simulation of dE/dx and ToF. Such a sample provides a reasonable description of all the kinematical variables associated with a given particle and the correlations among them. We then fit the composition of the particles produced in the vicinity of the B^- meson[6].

The *ToF* response for a particle depends on two observables: the momentum and the length traveled by the particle (also called arc-length). Consequently, the use of the *ToF* requires, in principle, a two-dimensional binning in momentum and arc-length. However, given the cylindrical geometry of the CDF detector, momentum and arc-length can be replaced by the momentum component in the transverse plane p_T and the pseudorapidity η[7] with no loss of information. Tab. 2 reports the total fraction of each particle type and Fig. 2 shows the p_T distributions resulting from the fit.

Combining the particle content and the covariance matrix estimated in each p_T bin, we are able to produce the *sPlot* of other kinematical variables for the whole Monte Carlo sample. Some examples are shown in Figs. 3-5: track η, the distance $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ between the particle and the B meson flight directions, the longitudinal component p_L^{rel} of the particle momentum with respect to the B meson flight direction (this variable is often used in the B flavor tagging algorithms).

A very good agreement is observed between the true distributions and the estimated ones, even in those cases where the plotted variable is strongly correlated to the p_T.

	π	K	p	e
Generated	0.828	0.112	0.0557	0.0039
Estimated	0.829 ± 0.002	0.111 ± 0.001	0.0557 ± 0.0006	0.0044 ± 0.0003

Table 2: Comparison between the true particle content in the Pythia Monte Carlo sample (Generated) and the fractions resulting from the fits (Estimated).

[6] This is a typical problem encountered during the development and the calibration of flavor tagging algorithms for the B meson.

[7] The pseudorapidity of a track is defined as $\eta = -\log(\tan(\frac{\theta}{2}))$, where θ is the polar angle.

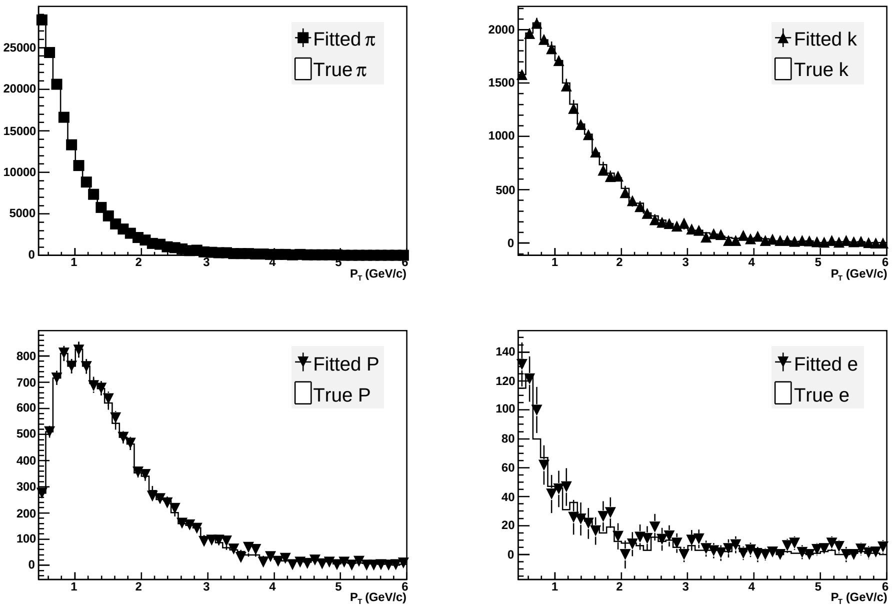

Figure 2: Overlay of the MC p_T distributions (filled histograms) and the result from the fit (dots).

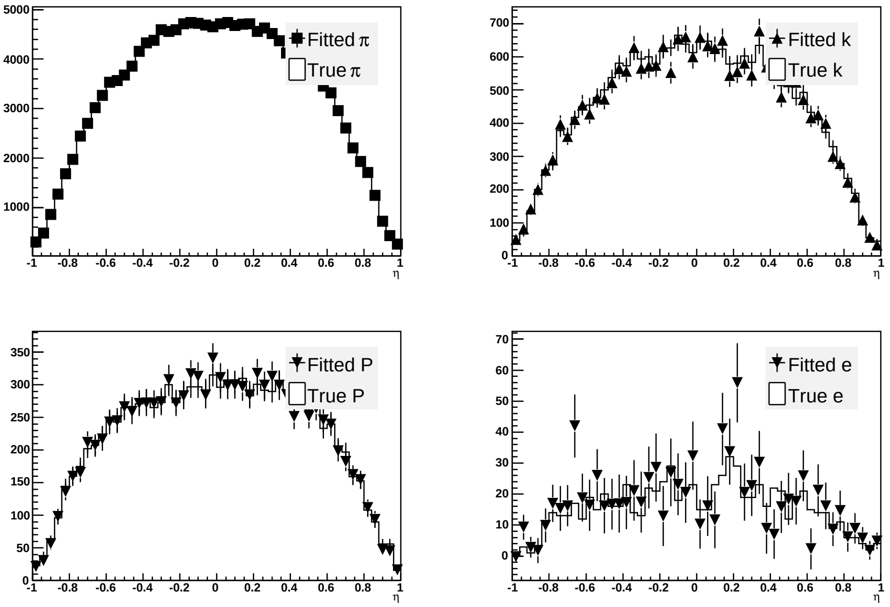

Figure 3: Pythia MC η distributions overlaid to the corresponding data *sPlot*.

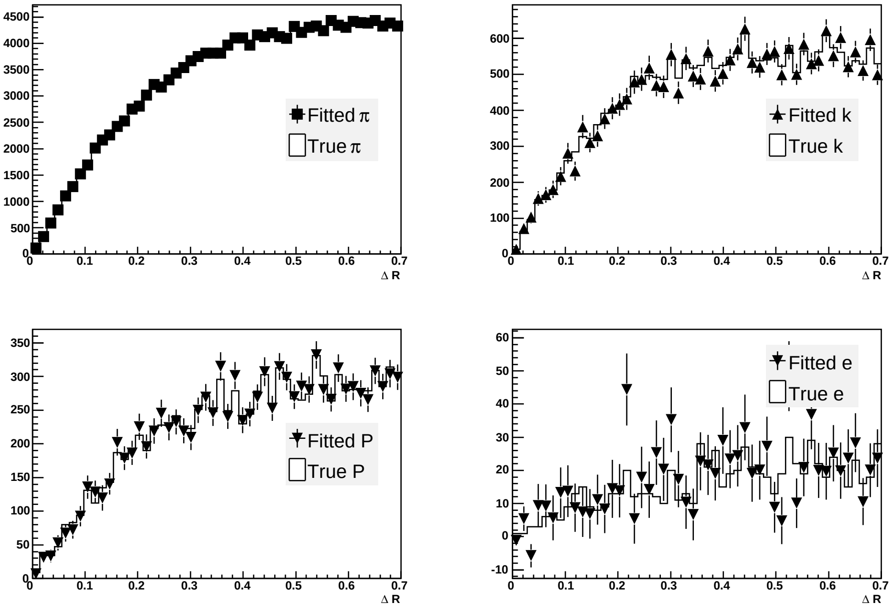

Figure 4: Pythia MC ΔR distributions overlaid to the corresponding data *sPlot*.

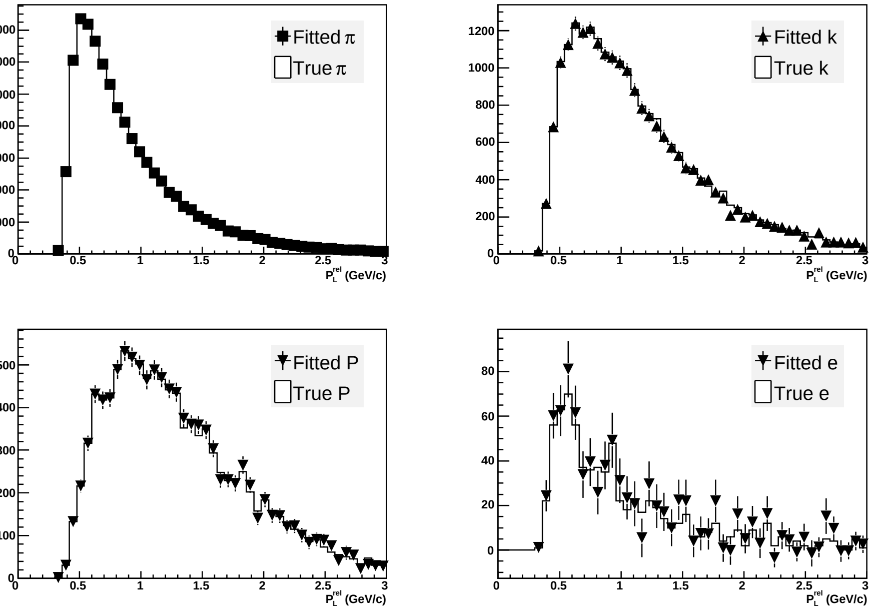

Figure 5: Pythia MC p_L^{rel} distributions overlaid to the corresponding data *sPlot*.

5 Conclusions and outlook

We have presented a method to estimate the distributions of kinematical variables of different particle species using information from *PID* detectors. This method is shown to work very well on Monte Carlo data allowing the determination of the fractional composition of a mixed sample of particle types with remarkable precision. We use a likelihood function that correctly contains the *pdf*'s of all the relevant event observables and therefore avoids a common mistake leading to strongly biased estimations. This approach looks very promising for the determination of distributions of different types of particles produced, for example, in conjunction with a B meson in a completely data-driven mode, without making any prior assumption on their functional forms.

6 Acknowledgments

We would like to thank the CDF Collaboration for providing the Monte Carlo samples used in this paper and for the very useful discussions and suggestions.

References

[1] CDF Collaboration, Phys. Rev. Lett. 97 (2006) 242003.

[2] G. Punzi, Comments on Likelihood Fits with Variable Resolution, PhyStat03, Stanford, California, USA, 2003 (arXiv:physics/0401045).

[3] P. Catastini and G. Punzi, Bias-Free Estimation in Multicomponent Maximum Likelihood Fits with Component-Dependent Templates, PhyStat05, Oxford, United Kingdom, 2005 (arXiv:physics/0605130).

[4] T. Sjöstrand *et al.*, Computer Phys. Commun. 135 (2001) 238. We use version 6.216.

[5] CDF Collaboration, Phys. Rev. D 71 (2005) 032001; CDF Collaboration, The CDF-II Techincal Design Report, FERMILAB-PUB-96-390-E, 1996.

[6] G. Cowan, Statistical Data Analysis, Oxford University Press, Oxford, 1998.

[7] J.M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, SIAM publications, Philadelphia, 2000.

[8] P.E. Condon and P.L. Cowell, Phys. Rev. D 9 (1974) 2558.

[9] M. Pivk and F.R. Le Diberder, Nucl. Instrum. Meth. A 555 (2005) 356 (arXiv:physics/0402083).